To determine the most appropriate multiple linear regression model for the anticipated county-level cancer death rate

Saurav Jaglan
School Of Computing
National College Of Ireland
Dulin, Ireland
x22105433@student.ncirl.ie

***Abstract*—The aim of this project is to determine which is the best multiple linear regression model to predict death in a cancer patient, this study is being performed on the cancer data set in CSV format, having socio-economic data of all the counties in the United States Of America.**

I. Introduction

Multiple Linear Regression is a machine learning technique, very well known and utilised for prediction or forecasting. This algorithm is said to be utilised when we are trying to determine the relationship between a continuous dependent variable and two or more independent variables.

II. Building The Machine Learning Model

A. Reading The Data-set and Descriptive Analysis

The first step prior to starting to work on any machine learning model is to identify in which format our data set is and thus we can use the appropriate function to read into our environment. On this occasion, our file is named "cancer.csv" and is in CSV format, so we will be utilising "read.csv()"function [1]. To proceed is important that we don't have null values present in the data set, null values can cause variation in our analysis, and sometime they may even result in different errors during the compilation of the code [2].

Statistics is a form of study where we perform analysis on the sample of the population instead of the population itself, so we will be splitting data into training data, which will be used to build our machine learning model and testing data, which later will provide the evidence of the efficiency of the model.

Once we have our training data we can get an insight into it, using the "summary()" function [3]. This will provide information about each column of the data set, the minimum and maximum values, and also the quarterlies.

When building a Multiple Linear Regression model, is best to check the correlation of independent variables with the dependent variable in this case "Death Rate".

No independent variables are showing a high correlation(between 0.9 and 1.0 c [4]) with the dependent variable, we can confirm this from Fig. 1. But there are columns which show a correlation between 0.4 to 0.52, such as: "Incidence Rate", "Percentage Bachelor Degree Age 25 Over", "Percentage Public Coverage Alone", "Percentage Public Coverage", "Percentage Private Coverage", "Poverty Percent".

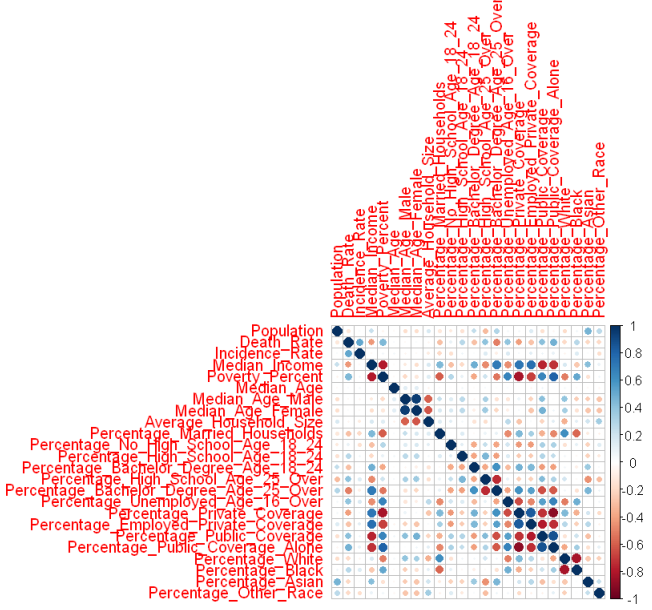

Fig. 1. Correlation Matrix

B. Model Building

The R function to build a machine learning model is "lm()" [5]. Even after finding the variables which are well correlated with the independent variable is hard to determine if those variables will be best suited to fit in the model for good accuracy of the model, so we will initiate a model with all the possible independent variables and check the summary of the model which will return us the different parameters to check the performance of our model Fig. 3.

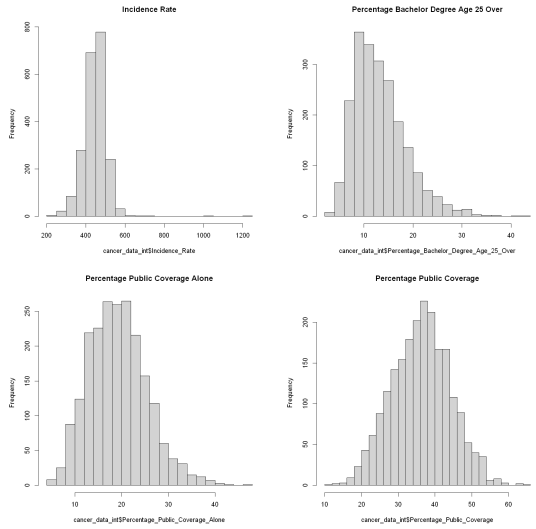

Fig. 2. Histograms

```
Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            1.959e+02  1.898e+01  10.317  < 2e-16
Population                            -1.903e-06  1.388e-06  -1.371 0.170580
Incidence_Rate                         2.026e-01  7.953e-03  25.475  < 2e-16
Median_Income                          5.939e-05  9.603e-05   0.618 0.536396
Poverty_Percent                        2.742e-01  1.753e-01   1.564 0.118031
Median_Age                             5.274e-03  9.384e-03   0.562 0.574120
Median_Age_Male                       -2.858e-01  2.500e-01  -1.143 0.253053
Median_Age_Female                     -9.181e-02  2.684e-01  -0.342 0.732303
Average_Household_Size                -1.846e+01  3.291e+00  -5.609 2.30e-08
Percentage_Married_Households         -1.886e-02  1.210e-01  -0.156 0.876159
Percentage_No_High_School_Age_18_24   -2.504e-02  6.880e-02  -0.364 0.715966
Percentage_High_School_Age_18_24       2.874e-01  5.935e-02   4.842 1.38e-06
Percentage_Bachelor_Degree_Age_18_24  -4.115e-02  1.305e-01  -0.315 0.752553
Percentage_High_School_Age_25_Over     1.914e-01  1.165e-01   1.643 0.100620
Percentage_Bachelor_Degree_Age_25_Over -1.303e+00  1.831e-01  -7.116 1.51e-12
Percentage_Unemployed_Age_16_Over      2.366e-01  1.957e-01   1.209 0.226750
Percentage_Private_Coverage           -8.027e-01  1.630e-01  -4.925 9.11e-07
Percentage_Employed_Private_Coverage   4.341e-01  1.229e-01   3.534 0.000419
Percentage_Public_Coverage            -2.846e-01  2.657e-01  -1.071 0.284170
Percentage_Public_Coverage_Alone       3.904e-01  3.342e-01   1.168 0.242923
Percentage_White                      -1.610e-01  6.712e-02  -2.399 0.016531
Percentage_Black                      -4.719e-02  6.343e-02  -0.744 0.457013
Percentage_Asian                      -2.246e-02  2.161e-01  -0.104 0.917237
Percentage_Other_Race                 -8.905e-01  1.456e-01  -6.117 1.13e-09

(Intercept)                           ***
Population
Incidence_Rate                        ***
Median_Income
Poverty_Percent
Median_Age
Median_Age_Male
Median_Age_Female
Average_Household_Size                ***
Percentage_Married_Households
Percentage_No_High_School_Age_18_24
Percentage_High_School_Age_18_24      ***
Percentage_Bachelor_Degree_Age_18_24
Percentage_High_School_Age_25_Over
Percentage_Bachelor_Degree_Age_25_Over ***
Percentage_Unemployed_Age_16_Over
Percentage_Private_Coverage           ***
Percentage_Employed_Private_Coverage  ***
Percentage_Public_Coverage
Percentage_Public_Coverage_Alone
Percentage_White                      *
Percentage_Black
Percentage_Asian
Percentage_Other_Race                 ***
```

Fig. 3. Summary

We will start optimizing our model by removing all the non-significant independent variables. The significance level is the possibility of refusing the null hypothesis when it is actually true. The significance codes indicate how much impact an independent variable is on the value to be predicted. A significance threshold of 0.001, for instance, means that there is less than a 0.1 per cent probability that the coefficient will equal zero and be considered inconsequential [6]. R language made it easy for us to recognise which values we should avoid including in our model, we can see the variable with three ”*” symbols, those are all significant independent variables, which we must fit into our model.

```
Residual standard error: 19.6 on 2130 degrees of freedom
Multiple R-squared:  0.5208,    Adjusted R-squared:  0.5193
F-statistic: 330.8 on 7 and 2130 DF,  p-value: < 2.2e-16
```

Fig. 4. Adjusted R Squared Value

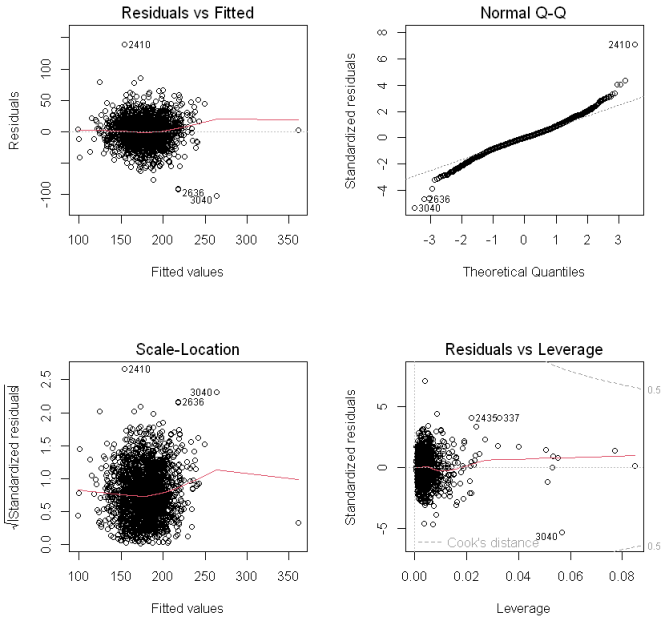

Fig. 5. Model Plot

C. Model Optimization

Once fit all the significant independent variables in the model. We can observe the value of Adjusted R Squared to be .51 Fig.4., for a model to be considered accurate, this value should be as high as possible, but it is also not possible to put faith in a model having an Adjusted R Squared value of 1, that’s simply because there is a possibility we included the dependent variable among the independent variables which in result will give us perfect machine learning model. So that model won’t be accurate and could not be used for prediction [7].

The next step is to verify if our model satisfies the Gauss Markov assumption, which can be easily done in R programming with the help of the graphical representation of our model.

1) Residual Plot: The difference between the dependent variable’s observed data and its fitted values, ŷ, constitutes the residual plot of a simple linear regression model [8].

$$Residual = y - \hat{y} \tag{1}$$

2) Q-Q Plot: Points on the Normal Q-Q plot demonstrate if the data set is uni-variately normal. The data points will sprawl on the 45-degree dotted line if the data is properly allocated. The points will be far away from the dotted line if the data points are not properly allocated or we can say distributed.

The Normal Q-Q plot is shown in the diagram below, where the relevant quantile values for the data set are represented

```
cancer_data_int <- subset(cancer_data_int, subset= !as.integer(rownames(cancer_data_int)) %in%
 c(2410, 3040, 337, 2435, 2636))
```

Fig. 6. Removing Outliers

```
Residual standard error: 18.5 on 2115 degrees of freedom
Multiple R-squared:  0.561,     Adjusted R-squared:  0.5596
F-statistic: 386.2 on 7 and 2115 DF,  p-value: < 2.2e-16
```

Fig. 7. Adjusted R Squared Value Model 1

on the y-axis and the quantile values for the standard normal distribution are plotted on the x-axis. The spots' proximity to the 45-degree reference line can be seen. At high levels of ozone exposure, the largest deviation from this line happens [9].

3) Spread-Location plot: Commonly named as the Spread-Location plot. This graphic demonstrates if residuals are distributed similarly across the predictor ranges. Here is how the assumption of equal variance can be tested (homoscedasticity). A horizontal line with evenly (randomly) spaced points is ideal [10].

4) Residual vs Leverage Plot: A particular sort of diagnostic graphic that enables us to pinpoint significant observations in a regression model is the residuals vs. leverage plot. The plot displays each observation from the data set as a separate point. Each point's leverage is displayed on the x-axis, and its standardized residual is displayed on the y-axis.

Leverage is the degree to which the regression model's coefficients would alter if a certain observation were omitted from the data set. The coefficients in the regression model are significantly influenced by observations with high leverage. The model's coefficients would change significantly if we removed these observations [11].

Outliers are values which can be extremely high or extremely low compared to the rest of the data. In all 4 plots we can observe the presence of outliers which possibly be a problem for the performance of our model, outliers handling is a very important step in model building. The presence of outliers can increase or decrease the performance of the model, the best practice is to remove outliers always. Once identified outliers we can remove those using the code in Fig.6.

Once Outliers are removed from our data set, we can proceed with building the model again using the same independent variables, to check if the performance of our model increases or decreases, then after that we have to also evaluate the 4 plots to verify the Gauss Markov assumption.

We can notice an increase in the value of Adjusted R Squared(from 0.5441 to 0.5543), the rise is not big but by this, we can be sure of if we keep removing outliers we can not only keep improving Gauss Markov assumption plots also we will be improving our model performance to forecast.

Once we are sure all outliers which might be problematic are removed we can observe our final model, and again verify the Gauss Markov assumption in Fig.7 and Fig 8. respectively.

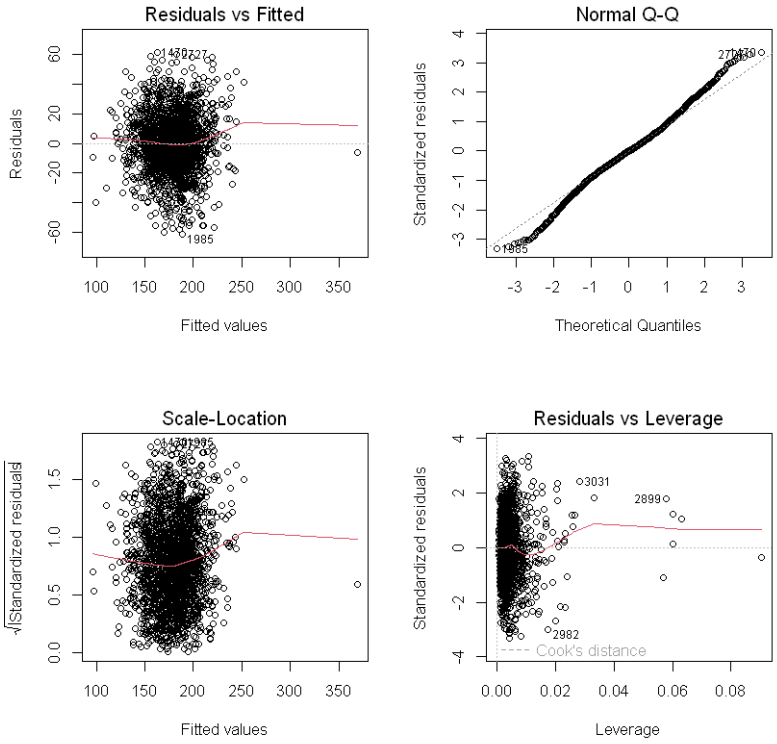

Fig. 8. Gauss Markov assumption plots

D. Use of OLSRR

OLSRR is a package present in the R programming language, which contains a set of functions which facilitates our model building [12], some of these are:

1. Collinearity diagnostics
2. Variable contribution assessment
3. Variable selection procedures

One function named "ols step all possible()"returns us all the possible models which we can build using the independent variables which we utilised to build our previous model i.e. model 1 for cancer death forecasting. We can observe in Fig.9. the graphical representation of all the models

Graph Fig.9 represents the plot of all the possible models against all the different evaluation methods, such as AIC (Akaike information criterion), Adjusted R Squared, Cp and R Squared. We don't use R Squared in multiple linear regression because its value raises as we add more independent variables, that's why an improved version of it, simply known as Adjusted R squared was calculated, The value of adjusted R Squared should be high as much as possible in this case our model containing all the independent variables with high significance value which we used in building model 1 has the highest Adjusted R Squared Value, now another evaluation we can look to is AIC, the AIC value should be lowest so again our model passes this evaluation as well. We can say the 1st model to detect deaths in cancer patients is a good model for prediction. But we can't be sure of it, we might be able

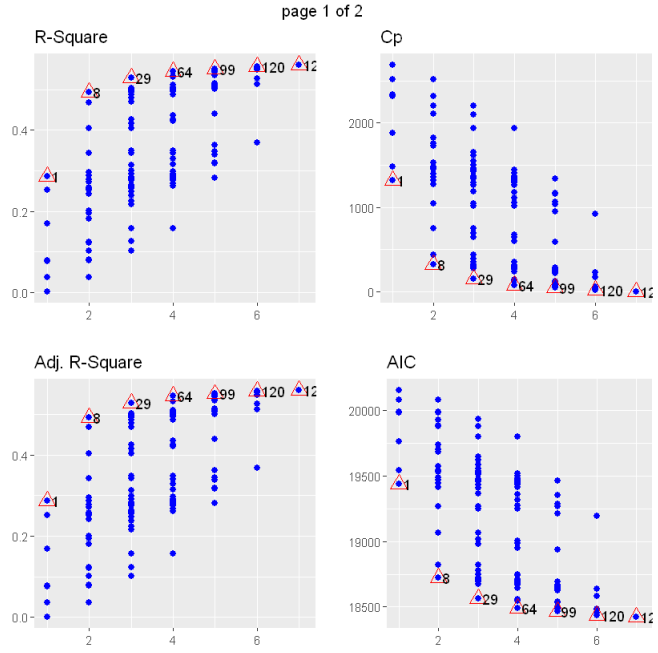

Fig. 9. OLSRR Plot

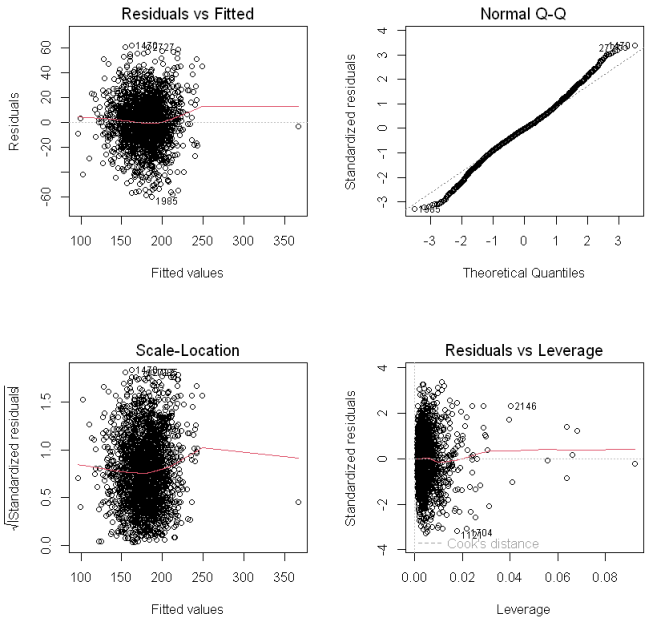

Fig. 11. 2nd Model Gauss Markov assumption Plot

```
Call:
lm(formula = Death_Rate ~ Incidence_Rate + Average_Household_Size +
    Percentage_High_School_Age_18_24 + Percentage_Bachelor_Degree_Age_25_Over +
    Percentage_Private_Coverage + Percentage_Employed_Private_Coverage *
    Percentage_Unemployed_Age_16_Over + Percentage_Other_Race,
    data = cancer_data_int)
```

Fig. 10. Independent Variables Transformation

to build a better one by bringing some modifications in the independent variables.

E. Transformation On Independent Variables

Sometimes even the model achieved using the independent variables present in the data set might not be the best model used for forecasting, sometime we may need to engineer these independent variables to create new independent variables. The aim is always to have a model with higher precision and low errors. The possible modification we can is multiplying the Percentage Employed Private Coverage and the Percentage of Unemployed Age 16 Over Fig.10.

This 2nd model's Adjusted R Squared value is 0.5639 which is higher compared to the 1st model's final value of 0.5596, so it is possible that this model might be a better choice, but to reach any conclusion firstly we need to verify if it verifies Gauss Markov assumption. The multiplication of two variables might result in originating outliers, so before reaching the final iteration of this model we might need to remove all the possible outliers, which can result in contradicting the Gauss Markov assumption.

The second model also satisfies the assumption, and all 4 plots show a good distribution of data but as we kept removing the outliers there was no impact on the value of Adjusted R squared, this simply means it is not fruitful to remove more

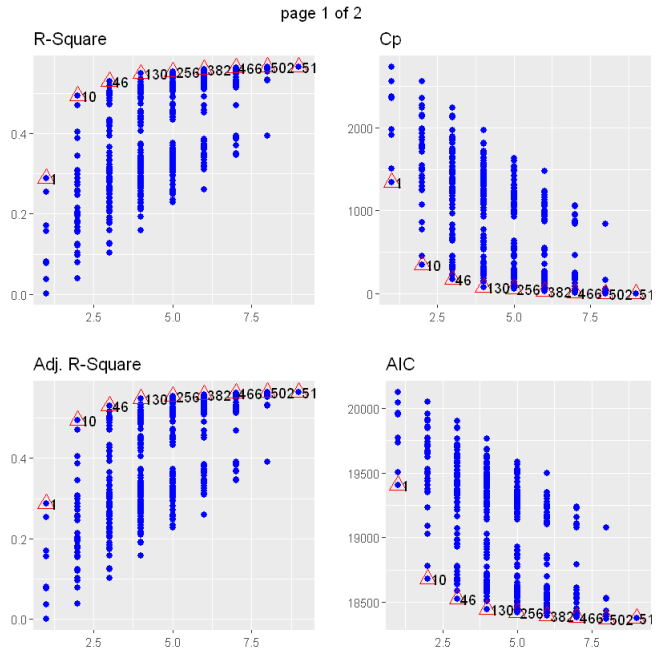

Fig. 12. 2nd Model OLSRR Plot

outliers from the data, as they won't affect the performance of the model. We can assume this is the best variation of the model we achieved after the modification in independent variables. To verify this we can again perform a check using the OLSRR function "ols step all possible()".

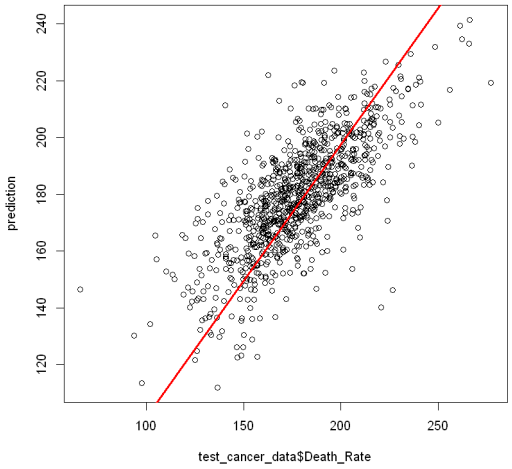

Fig. 13. Correlation Plot for 1st Model

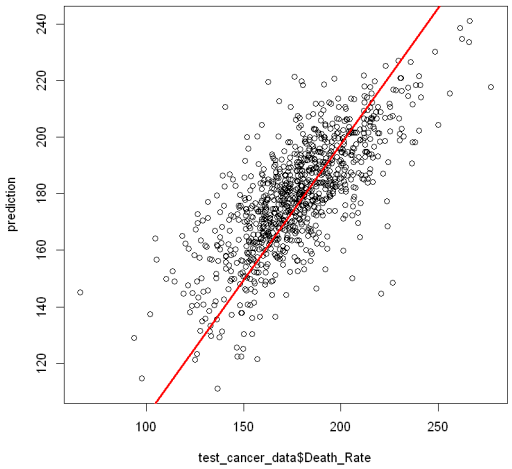

Fig. 14. Correlation Plot for 2nd Model

F. Testing Data and Model Selection

Now that we have two models, which both satisfy the Gauss Markov assumption and are free from outliers value, we need to test these models so we can identify which among the two is best suited for forecasting, for prediction use the ”predict()” function [13]. Once predicted we store the predicted values in a variable and plot a scatter plot graph to check the correlation between actual values and predicted values. This is done with both models, hence we can observe Fig.13.

G. Conclusion

To conclude we can identify that the first model has a higher correlation with the actual values, so the 1st model is the best fit to predict the death rate because of cancer, after evaluating the model using different evaluation methods and having an Adjusted R Squared value of 0.55 can be further improved but the value according to my assumption won't exceed 0.57 or 0.58.

REFERENCES

[1] https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/
[2] https://statisticsglobe.com/r-is-null-function/
[3] https://www.educative.io/answers/what-is-the-summarize-method-in-r
[4] https://www.andrews.edu/ calkins/math/edrm611/edrm05.htm
[5] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm
[6] https://infocenter.informationbuilders.com/wf80/topic/pubdocs/RStat16/source/topic41.h
[7] https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp
[8] https://www.r-tutor.com/elementary-statistics/simple-linear-regression/residual-plot
[9] https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/normal-qq-plot-and-general-qq-plot.htm
[10] https://data.library.virginia.edu/diagnostic-plots/
[11] https://www.statology.org/residuals-vs-leverage-plot/
[12] https://cran.r-project.org/web/packages/olsrr/vignettes/intro.html
[13] https://www.digitalocean.com/community/tutorials/predict-function-in-r